DOCUMENT RESUME

ED 209 289 TM 810 790

AUTHOR Robertson, Gary J.
TITLE Assessing Reliability of Criterion-Referenced Instruments.
PUB DATE 30 Apr 81
NOTE 19p.; Paper presented at the Annual Meeting of the International Reading Association (26th, New Orleans, LA, April 30, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Reading Tests; Scores; *Test Reliability
IDENTIFIERS *Domain Referenced Tests; Estimation; *Reliability Formulas

ABSTRACT

Some fundamental concepts of criterion referenced test (CRT) reliability are highlighted. Emphasis is given to the procedures for determining reliability of scores for individual pupils because this is an area requiring increased awareness by classroom teachers and practitioners. Reliability issues encountered in the evaluation of instructional programs are not considered, however. After some preliminary remarks about reliability, methods of estimating CRT reliability are discussed as are additional considerations affecting CRT reliability estimates. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. J. Robertson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED209289

TM 810 790

Assessing Reliability of Criterion-Referenced Instruments

Gary J. Robertson, Ph. D.
American Guidance Service

Developers and purveyors of criterion-referenced tests offer scant technical information of the sort considered so essential in the evaluation of norm-referenced tests. Perhaps freedom from the rigors of norm sample definition and the standardization process have been interpreted by criterion-referenced test (CRT) developers as license to avoid any sort of prepublication research to establish relevant properties of the measuring instrument(s). Lack of uniform developmental guidelines for CRTs such as the AERA/NCME Standards for Educational and Psychological Tests which apply to norm-referenced tests (NRTs) has impeded the highly desirable practice of establishing such fundamental psychometric qualities as test reliability and validity. This observation is underscored by the author's recent review of six CRT reading assessment systems developed for the primary and elementary grades.[1] In only one of these publications was any technical data on test reliability presented; the other five publications contained no technical data of any sort. Consumers are forced to accept on faith the technical adequacy of these measures. Such practices seem unlikely to change until a compelling set of standards and significant competition in the marketplace effect a change in present CRT test development practices.

---

[1] These are not cited by name here because the purpose of this article is not to critique specific tests.

Paper presented to the 26th annual convention of the International Reading Association, New Orleans, Louisiana, April 30, 1981. The author wishes to thank the following persons for their comments and suggestions: Dr. Gerald S. Hanna, Kansas State University; Dr. Anthony Nitko, University of Pittsburgh; Dr. Leo Schell, Kansas State University.

In all fairness, it is important to mention that the technology of CRT development is still in an embryonic state. It is only within the past 10 years that much of the underlying theory has been developed, and much still remains to be done. Classical test theory, by way of contrast, has evolved over more than 60 years and represents a distillation of principles that have withstood repeated empirical testing and refinement. Perhaps CRT technology will be in a similar state 50 years hence.

This paper was written to highlight some fundamental concepts of CRT reliability. Emphasis is given to the procedures for determining reliability of scores for individual pupils because classroom teachers and practitioners find this to be the area of greatest need. Reliability issues encountered in the evaluation of instructional programs are not addressed in this paper.

Some Preliminary Remarks About Reliability

Test reliability refers to the consistency with which a measuring instrument performs its intended function. Estimates of reliability for NRTs have typically included (1) consistency of performance on a sample of test items administered at a single sitting (internal consistency) or (2) consistency of performance on the same or parallel test forms typically administered on two separate occasions but occasionally administered at a single sitting. Both classes of estimation procedures employ correlational methods whose magnitude is dependent upon the degree of variability, or spread of scores,

present in a particular reference group of interest. NRTs are constructed to maximize the variability among examinees, while CRTs are constructed to sample a carefully specified content domain with little or no attention given to item or score variability (Linn, 1979). Thus, the presence or absence of test score variability has been cited as a fundamental distinguishing characteristic between norm-referenced and criterion-referenced measurement (Popham and Husek, 1969, p. 3).

In general, it seems safe to conclude that procedures typically used to estimate NRT reliability are less valuable with CRTs. Alternative methods appropriate for evaluating CRT reliability will be presented in this paper. Criterion-referenced reliability estimates, like norm-referenced reliability estimates, become meaningful only when viewed in conjunction with the inferences to be made from the test results. For norm-referenced tests, item selection is guided by content considerations, not by item difficulty or discrimination indices. The absence of high classical reliability coefficients is not of great concern to CRT developers because such information is viewed as largely irrelevant to the stated purpose of such measures. According to Linn (1979), test variability is, nevertheless, almost always present to some degree in CRTs; and "the use of traditional indices need pose no real danger if not allowed to distort content coverage. If traditional reliability indices are high, this is informative. If they are low, they are less informative, especially if ... the variability of test scores or ... items is small" (p. 94). As with traditional norm-referenced

reliability estimates, criterion-referenced reliability estimates become meaningful only when the intended uses and inferences that are to be made from the test scores are considered. These issues will be considered briefly in the next section.

Members of the audience may have been wondering why there is "so much fuss" by measurement specialists about reliability. Teachers and practitioners are, after all, responsible for educating pupils and have neither the time nor inclination to pursue "esoteric measurement topics" such as CRT test reliability. It is well to keep in mind that diagnostic-prescriptive reading instructional systems with CRT components are rapidly finding their way into schools. Developers of these devices have presented little evidence of their technical characteristics, especially their reliability. Such omissions raise serious doubts about the meaning of scores on these devices because test reliability is a necessary, if not sufficient, requirement for validity; moreover, the success of any measuring instrument ultimately rests on its validity for one or more stated purposes.

Methods of Estimating Criterion-Referenced Reliability

Reliability estimation procedures for criterion-referenced measures may be classified as indices of (1) consistency of classifications, (2) consistency of scores, or (3) domain score estimation. Each of these estimation procedures will be described briefly below. It should be mentioned here that the term agreement index may be more appropriate for the indices described below than reliability coefficient (Berk, 1979, p. 324).

Consistency of Classification.[2] Criterion-referenced tests are frequently used to classify pupils into mastery and nonmastery categories. A "cutting" score is typically set[3] with all pupils scoring at or above that point termed "masters" and all scoring below that point termed "nonmasters." A logical question arising from this practice relates to the extent of agreement in making these classifications using data from a second testing with the same test or a parallel form of the test. Decision consistency is usually measured by the percentage of agreement in making the mastery-nonmastery classification on first and second testings. Results are typically presented as shown in Table 1 (a 2 x 2 contingency table) for a fictitious set of data. Note that the percentage of consistent classifications is obtained by summing cells a and d. Thus, a rather impressive 87% of the pupils were consistently classified when results from the two testings were considered jointly.

A second index of the extent of agreement on repeated measures is Cohen's (1960) Kappa coefficient (Kappa). Kappa measures the test's contribution to the overall proportion of consistent classifications by taking into account agreement that could be expected to occur by chance alone (as evidenced by the mastery-nonmastery characteristics of a particular group of pupils when tested on two separate occasions).

---

[2]Technically, these indices are termed threshold loss functions.

[3]Details of procedures used to establish cutting scores are beyond the scope of this paper, but interested readers may wish to consult Hambleton's review in Berk (1980a).

Kappa is calculated by subtracting from the observed percentage of agreement that expected to occur by chance. For the data in Table 1, the Kappa coefficient is .68.[4] Thus, the agreement seems less good after elimination of the effects of chance agreement. It should be noted, however, that certain problems with the Kappa coefficient have resulted in the recommendation that the uncorrected percentage of agreement statistic be used instead of Kappa. Problems stem from instability of Kappa for short tests typically used by classroom teachers, constraints upon the range of Kappa imposed by the marginal frequencies in the contingency table, and sensitivity of Kappa to the cutting score, with higher values of Kappa tending to occur when the cutting score is set near the mean and lower values tending to occur when the cutting score falls in the tails of the test score distribution (Berk, 1980, pp. 332-333).

The data summarized in Table 1 are not much help to the classroom teacher interested in more detailed information about the degree of mastery or nonmastery of class members because all degress of mastery

---

[4] Obtained by using the formula

$$\text{Kappa} = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the observed percentage of agreement, and P_c is the percentage of the observed agreement expected to occur by chance. In Table 1, P_o is 87 percent and

$$P_c = (.26)(.31) + (.74)(.69) = .59.$$

$$\text{Kappa} = \frac{.87 - .59}{1 - .59} = .68$$

and all degrees of nonmastery are treated the same. Teachers are often interested in identifying those pupils of lowest nonmastery status in order to give special remedial aid. For this reason, the data presentation scheme shown in Table 2 is preferable for use in situations where more specific information about mastery score levels is needed. The percentage of agreement (87%) is identical in Tables 1 and 2 although the Kappa coefficient changes in Table 2.

Teachers and school personnel are urged to conduct such studies locally and tabulate the test results as shown in either Table 1 or Table 2. If a parallel test form is unavailable, then the same test may be readministered a few days after the first test. The data analysis procedures are simple and straightforward and will eventually provide a framework with which the performance of locally developed or teacher-made tests can be evaluated. Special procedures are available to estimate classification consistency where the administration of only one form of a test is possible (see Berk, 1980a, pp. 131-150 for a summary of approaches devised by Huynh, Subkoviak, and Marshall & Hertel). The Hambleton and Novick two-administration method is recommended, however, for tests developed locally by classroom teachers.

Consistency of Scores. Reliability indices of this sort are based on the squared deviations of individual scores from the cutting score.[5]

---

[5]Technically, these indices are termed squared error loss functions.

As such, these estimates are much more sensitive to the degree of mastery or nonmastery along the score scale. These estimates reflect all measurement errors present, not just those leading to misclassifications. The two reliability indices of this type are described in detail by Berk (1980b, pp. 335-339). One appealing aspect of these methods is their use of all test score data, not merely the master-nonmaster categorization. Hence, these methods seem appropriate where teachers use tests to place students in instructional treatments. An added advantage is that only one administration is required.[6] A chird benefit from using Brennan's statistic is that aspects of generalizability theory can be applied to criterion-referenced tests to estimate variance components of interest in a particular situation (Berk, 1980b, p. 339).

Domain Score Estimation. Domain score estimation procedures are analogous to calculating the standard error of measurement in classical test theory. Thus, domain score estimation is markedly different from the estimation of indices of agreement characteristic of the two preceding classes of estimation procedures. Essentially, domain score estimation consists of computing standard error estimates which can subsequently be used to construct confidence intervals around an individual's score, expressed as the observed proportion correct. In practice, interval estimates can be derived either for an individual's percent correct score or for the cutting

---

[6]Although an important consideration relates to whether there is an underlying assumption of classically parallel or randomly parallel test forms.

score used to define the mastery-nonmastery classifications. Defining the cutting score as a band of scores rather than an exact point is appealing because it takes account of measurement error in establishing the cutting score. A problem with domain score estimates is that they are sensitive to test length, composition of the group tested, and form of score distribution. Estimates of percentage of mastery scores based on less than 10 items are especially unstable and should be used cautiously within this context. It is recommended that this approach to domain score estimation will have the greatest utility if errors of measurement are related to the cutting score (Berk, 1980b, p. 345).

Additional Considerations Affecting Reliablity Estimates for Criterion-Referenced Tests

In addition to the technical characteristics of the methods of estimating CRT reliability outlined above, consideration must also be given to the structural properties of the content domain, the nature of the examinees, and the nature of the measuring device. These must all be viewed within a specific decision-making context.

Properties of the Content Domain. A recent paper by Nitko (1980) outlines his proposed scheme for classifying the many varieties of educational achievement tests, primarily criterion-referenced, but also norm-referenced measures. Although the scope of the present paper precludes a detailed examination of Nitko's classification scheme, his proposed structure provides interesting insights into

the types of behaviors measured by various reading tests. One can see that selected measures of reading ability are classified in different definitional categories, with concomitant implied differences in the type of reliability indices which might be appropriate. It should be made clear that Nitko does not link his taxonomy to specific methods of reliability determination. Such speculations are the responsibility of the author and imply no endorsement by Nitko.

Essentially, Nitko classifies CRTs into two basic groups:

(1) Well-Defined and Ordered Domains

(2) Well-Defined and Unordered Domains

Tests are categorized in Category 1 if there is a logical or empirical basis for ordering the behaviors constituting the domain. Bases for ordering include an underlying developmental continuum, well-defined and accepted proficiency levels, an empirically defined latent trait, or judged aesthetic or social quality of an examinee's product or performance. Tests are classified in Category 2 if the domain is narrow and comprised of homogeneous items, or if there is no logical or empirical basis for ordering behaviors. Table 3 shows the diversity with which some selected measures of reading skills are classified by using Nitko's scheme.[7] Implications for different types of reliability indices most appropriately reported for stated uses or purposes are also given in Table 3. The tasks

[7] Reading was selected for this analysis to take into account the interests of members of the audience. Other content areas could have been selected.

listed in Table 3 are not exhaustive of all reading tasks, nor are the reliability indices suggested in all cases the only ones appropriate for the skills assessments listed.

As seen in Table 3, the extent to which the domain is ordered or well-defined, together with the use to be made of the results, influences the type of reliability estimate that is most appropriate. Estimation of domain scores, one of the three types of CRT reliability estimation procedures, is possible only for well-defined, ordered domains.

Table 3 also illustrates that both CRT and NRT tasks are needed to obtain the various skills assessments of most use to classroom teachers. A good reading skills testing program might well require both NRTs and CRTs to obtain the type of information needed. Neither CRTs nor NRTs alone provide every type of information needed in the assessment of reading skills. Moreover, classical reliability estimates may be appropriately used in well-defined and ordered domains characterized primarily by criterion-referenced measures. The determining factor in selecting the type of reliability estimate is the intended use of the test.

Reliability, especially the reliability of difference scores, also depends upon the extent to which a content domain can be fractionated into discrete, meaningful subbehaviors. Research by Drahozal and Hanna (1978) has raised serious questions about the legitimacy of

such attempts for the reading comprehension subskills of pupils in grades 3 through 9. For example, reliable differentiation among literal, translation, and inferential reading subscores was not supported when intercorrelations among subscores in these three areas were corrected for attenuation, or unreliability. Schell (1979) notes that reading instruction invariably follows a spiral curriculum: skills are taught, practiced, and reviewed at one level, and reintroduced, retaught, and practiced again at subsequent levels. There is apparently no such phenomenon as all-or-none mastery of a particular reading comprehension skill as is characteristic, for example, of certain mathematical skills. To derive and report separate reading comprehension subscores with either norm-referenced or criterion-referenced interpretation is an egregious practice all too characteristic of certain reading assessment systems currently in use in the schools.

The above examples were given to emphasize that the nature of the content domain has important implications for the type of reliability estimates that are sensible, logically appropriate, and permissible. Educators must constantly seek ways of becoming cognizant of such implications and limitations within the context of their own test use.

Nature of the Examinees. CRT reliability is dependent, as is true of classical norm-referenced reliability estimates, on pupil characteristics such as age and maturational factors, amount of

instruction, achievement motivation, and specific situational circumstances affecting a particular test. These sources of variation are important and are accounted for by some, but not all, CRT reliability estimation procedures.

Nature of the Measuring Device. In general, longer tests contain more extensive samples of behavior and, other things being equal, yield higher reliability estimates. A test with one or two items per domain or behavioral objective typically yields far less reliable estimates than a test with 15 items per domain or objective. Related to test length is the matter of guessing. With multiple choice and true-false tests, it is important to establish criterion, mastery, or cutting scores well above the chance score level that could be achieved by guessing. Test users are reminded, in particular, that domain score estimation procedures are especially sensitive to test length and should not be used for tests (or clusters of items) consisting of less than 10 items. For decision-consistency reliability, the Kappa coefficient of agreement is particularly unsuitable for short tests of 10 to 20 items. Uncorrected percentage of agreement statistics are relatively stable for subtests containing less than 10 items with little variability in scores. The Kappa coefficient is also sensitive to the mean, and assumes its highest values when the cutting score is closest to the mean. By way of contrast, percentages of agreement assume higher values when cutting scores fall in tails of the score distribution and some distance from the mean. In summary, test length, mean score, cutting score,

and variability of the scores are all important considerations in the selection of an appropriate CRT reliability index.

Concluding Statement

Various aspects of CRT reliability were surveyed briefly in this paper with the intended purpose of increasing the awareness of practitioners who deal on a day-to-day basis with a variety of criterion-referenced reading assessment devices now being purchased by schools. Lack of adequate technical information for such materials is the rule, rather than the exception.

It is hoped that consumers will be more demanding of the developers of these materials and that this paper will stimulate discussion among practitioners and assist them in bringing about some needed reforms.

References

Beck, R. A. Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press, 1980. (a)

Beck, R. A. A consumers' guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-349. (b)

Drahozal, E. C., & Hanna, G. S. Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, February 1978, 416-420.

Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington: National Council on Measurement in Education, 1979.

Nitko, A. F. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.

Popham, W.J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Schell, L. Criterion-referenced reading tests: Selected cautionary notes. Reading World, October 1979, 57-62.

Table 1

Decision-Consistency Reliability

For a Fictitious Data Set, Forms 1 and 2,

Administered on Successive Days

(N = 100)

First Testing (Form 1)	Second Testing (Form 2)		Total
	Masters (≥80% correct)	Nonmasters (<80% correct)	
Masters (≥80% correct)	22% (a)	4% (b)	26%
Nonmasters (<80% correct)	9% (c)	65% (d)	74%
Total	31%	69%	100%

Percentage of agreement = 87

Kappa coefficient: .68

Table 2

Decision-Consistency Reliability for a Fictitious Data Set Differentiated by Mastery Score Levels, Forms 1 and 2, Administered on Successive Days

(N = 100)

Mastery Score Level	First Testing (Form 1) %	Second Testing (Form 2) % with Identical Status
Masters		
≥ 80% correct	26	22
Nonmasters		
70-79% correct	13	11
60-69% correct	20	19
50-59% correct	23	20
40-49% correct	9	6
30-39% correct	5	4
29% & below correct	4	5
Total Nonmasters	74	65
Total Group	100	87

Percentage of agreement = 87

Kappa coefficient: .84

Table 3

Classification and Type of Reliability for Some Intended Uses of Selected Reading Skills Tasks

Reading Skills Task	Type of Domain[a]	Intended Use	Appropriate Reliability
1) Reading rate	Well defined/ordered: degree of proficiency	1a) Estimate of reading rate 1b) Rank at grade level	1a) Domain score estimate (CRT) 1b) Classical reliability estimates
2) Functional reading level (Degrees of reading power)	Well defined/ordered: difficulty or complexity	Estimate of functional reading grade level	Domain score estimate (CRT)
3) Decoding: beginning consonant blends	Well defined/unordered: delineation by stimulus properties	Master of content	Decision-consistency estimate (CRT)
4) Comprehension of scholastic reading material	Well defined/unordered: delineation by construct	Comprehension skills compared with grade reference group	Classical reliability estimates

[a]Outlined in Anthony J. Nitko, "Distinguishing the Many Varieties of Criterion-Referenced Tests." Review of Educational Research, 1980, 50.